Future Projections of the Large Scale Meteorology Associated with California Heat Waves in CMIP5 Models

Erool Palipane[1] and Richard Grotjahn[1*]

[1]*Department of Land, Air and Water Resources, University of California, Davis, CA, 95616, USA*

January 2018

Submission to Journal of Geophysical Research: Atmospheres

How to cite:

---

* *Corresponding author address*:
Atmospheric Science Program, One Shields Ave., Dept. of L.A.W.R. University of California Davis, Davis CA USA 95616
Corresponding author email: grotjahn@ucdavis.edu

Key Points (140 characters)

- The synoptic pattern for California Central Valley heat waves does not change in frequency or intensity in the future in climate models.
- Heat waves are much more frequent and predominantly of one type when using historical thresholds due to the change in the climate ‘mean’.
- A multi-model average has 4x as many heat waves, lasting 2x as long, with 1.5x the 20-year return value relative to historical values.

Abstract

Previous work showed that climate models capture historical large-scale meteorological patterns (LSMPs) associated with California Central Valley (CCV) heat waves including both ways these heat waves form. This work examines what models predict under the RCP4.5 and RCP8.5 scenarios. Model performance varies, so a multi-model average weights each model based on its historical performance in four parameters. An LSMP index (LSMPi) is defined using upper atmosphere variables that best capture dates of past extreme surface temperature maxima. LSMPi correlates well with all values of CCV surface maximum temperature. LSMPi distributions in future simulations shift ~0.6 standard deviations higher between 1961-2000 and 2061-2100 for RCP 8.5 data. Based on the *historical* climatology, future scenarios show a large increase in the frequency and duration of heat waves in every model. Four times as many heat waves occur and their median duration doubles, using historical thresholds. Of the two ways heat waves form, type 1 has similar frequency in the future. But, type 2 becomes much more common because type 2 has a preexisting hot anomaly in Southwestern Canada, much like the historical to future climatological change in that region (a “global warming” signal). The 20-year return value anomaly increases by 30-40%. The average of the 50 hottest temperatures increases 3.5-6K depending on the scenario. When extreme values are defined using the *future* climatology, the models and their average have no consistent increase or decrease of distribution properties such as: shape, scale, and return values of the extremes compared to historical values.

Index Terms (five or less)

3337 Global climate models (1626, 4928)

3305 Climate change and variability 4313 Extreme events (1817, 3235) 3364 Synoptic-scale meteorology 0429 Climate Dynamics. ~~1616 Climate variability, 4317 Precursors~~

Keywords (six or less): future heat waves, California heat waves, large-scale meteorological patterns simulation, future summer variability, climate model simulations of heat waves

1. Introduction

The California Central Valley (CCV) is the most agriculturally productive region in the world, and extreme heat is a major concern during the summer months of June through September. Prior work discusses how large-scale meteorological patterns (LSMPs) are associated with CCV heat waves (Grotjahn & Faure, 2008; Grotjahn, 2011, 2013), that CCV heat waves can form by two ways (Lee & Grotjahn, 2016; hereafter LG2016), and how the climate models vary in their ability to create simulated heat waves in historical conditions (Grotjahn & Lee, 2016; hereafter GL2016). This paper, applies the LSMP context to identify and understand possible future changes in CCV extreme heat events during summer. Our specific questions include: will events occur more often than in the past? Will events become more severe? Will events last longer? Will changes from historical to future climate be due mainly to a shift in the climatological conditions (a 'global warming signal') or in the LSMP properties?

Coumou et al. (2013), Perkins et al. (2012), Russo et al. (2014) and others have discussed the effects global warming may have on local heat event characteristics in the mid-latitudes. Other studies explore possible physical mechanisms behind the changes in these heat wave events. These mechanisms include: changes in sub-seasonal atmospheric variability (Teng et al., 2013), variations in the quasi-stationary waves (Screen et al., 2014; Petoukhov et al., 2013) and weakening of the boreal storm tracks in the summer (Lehmann et al., 2014). Understanding the physical mechanism is key to understanding and attributing the changes to the global warming signal.

Grotjahn (2011) and Horton et al. (2016) discuss heat waves synoptics: subsidence causing warming of the air from adiabatic compression and clear skies to support radiant heating, and advection of warm air. Also, Grotjahn (2011), Lau et al. (2012), and Grotjahn et al. (2016) discuss how offshore winds occur with severe heat waves in California. Finally, Grotjahn (2011) finds the largest temperature anomalies are just offshore (at 850 hPa), helping to set up the low level pressure gradient force that opposes a cooling sea breeze and the subsidence lowers the subsidence inversion leaving a shallow surface layer to warm by solar radiation. Large scale features associated with California heat waves (i.e. LSMPs) are resolved by climate models and provide a context to examine different models predictions of future CCV heat waves.

Grotjahn (2011) developed a LSMP index based on upper air data that matches well the surface heat wave temperatures over the CCV. Grotjahn (2013) shows how well the CCSM4 model simulates the LSMPs and other properties of heat waves compared to reanalysis data. LG2016 and GL2016 use a cluster analysis to sort CCV heat waves into two types based on LSMPs leading up to heat events. One cluster ("type 1") has cold anomalies prevailing over the NW US and western Canada several days before CCV heat event onset and the CCV heat wave develops quickly in the day before onset. The other cluster ("type 2") has a preexisting hot anomaly over SW Canada for several days prior to CCV heat onset, then a southwestward extension of the hot anomaly initiates the CCV heat wave.

This work builds on our previous work to answer those questions above. We consider 13 different CMIP5 models simulations of the RCP4.5 and RCP8.5 scenarios. We improve the diagnostics of the LSMPs from our prior work and apply those diagnostics to estimate how the extremes are going to change in the future, including the two cluster types. We develop a simple multi-model average based on each model's historical performance.

The next section describes the data and methods developed to understand the future changes in CCV heat waves. The third section describes the main results and section four describes the main conclusions.

2. Data and Methodology

2.1. Data Used for the Analysis

NOAA climate data at 17 stations in the CCV are used to identify historical extreme heat event information, principally the frequency of events.

The NCEP-NCAR reanalysis (Kalnay et al., 1996, hereafter NNRA1) are used for: the formulation of the LSMPs via composites, the development of an improved LSMP index, distinguishing the two cluster types of heat waves, and verification and comparison with corresponding quantities in the model data. The NNRA1 data are from the 40-year 1971-2010 period. ERA-interim (Dee et al., 2011) data are used to cross-check the NNRA1 results and the results in the overlapping period are essentially the same. Hence the NNRA1 reanalysis is used because it has a longer record of data and hence includes more extreme heat wave events.

CMIP5 model data from historical, RCP4.5, and RCP8.5 simulations are studied. Model historical data are from the 40 simulated years 1961-2000; the RCP simulations are for 2061-2100. Climate model simulations are not weather forecasts, so a specific date has only accidental similarity between models and the reanalysis. So, the model and reanalysis periods are offset to take advantage of better upper air observations at later times while the historical simulations end before 2010. Some models have parallel simulations (ensemble runs) with the sub-daily upper air data we need archived; as available, the data from ensemble runs were included. Table S1 and the GL2016 supplementary information give a description of the model data used and how many grid points for a specific model are considered to lie within the CCV.

The zonal wind anomaly (Ua), the meridional wind anomaly (Va) and the temperature anomaly (Ta) were examined at these levels: 250hPa, 500hPa and 850hPa at every 6hr snapshot time (i.e at 0hr, 06hr, 12hr, 18hr). The anomalies are with respect to corresponding long term daily mean values (LTDM). The LTDM values for each of these variables are found by the methodology described in LG2016. To summarize: corresponding days of the year are averaged to create an initial LTDM at each location; since the initial LTDM has sizeable day to day variation on a 40 year average, the data are Fourier transformed and the first five harmonics used to construct the smooth, final LTDM. Daily anomalies at each location are constructed by subtracting the corresponding daily final LTDM values.

For surface maximum temperatures over the CCV, an additional step is made to normalize the daily anomalies by the long term mean seasonal average standard deviation at each location. The resulting normalized anomalies make values at different locations inter-comparable. These normalized anomalies are labeled 'Tnamax' here. The spatial average of the Tnamax values over the CCV for each day is labeled '**avTnamax'.** Our prior work uses this methodology. Also, as in Grotjahn (2011), the event onset is always at 12 GMT.

2.2. Definitions of a Heat Event

We use a similar methodology as used by GL2016 for the identification of CCV heat waves in the CMIP5 models. We treat the grid points in the CCV region in a similar way that was done for station data. Daily maximum surface temperatures at the CCV grid points are saved for each model. At least ½ of the CCV grid points must reach or exceed a threshold for the date to qualify as a heat event. The threshold is the value of the 95th percentile of Tnamax at that location. Tnamax is used since the dimensional value of the threshold (the 95th percentile) and its variance varies between CCV locations over the summer. One difference from GL2016 is that instead of using 1972-2005, the period here is 1961-2000. So the models use 40 years like the reanalysis but start 10 years earlier since historical model simulations end in 2005.

The future scenarios use data from 2061-2100. We define heat waves in the RCP4.5 and RCP8.5 data in two different ways. One way uses the same threshold values of surface avTnamax as calculated in the historical simulations to define a future heat wave. The second way uses the 95th percentile from the normalized temperature values from the future time period calculated for each respective RCP scenario.

The following label conventions designate how the threshold is defined when choosing candidate heat waves.

1. Heat waves from CMIP5 historical runs use the threshold based on the model's historical data - CMIP5_Hh
2. Heat waves from the CMIP5 RCP runs that use the threshold based on the model's historical data – CMIP5_Fh (RCP45) and CMIP5_Fh (RCP85) for the RCP 4.5 and 8.5 scenarios respectively.
3. Heat waves from the CMIP5 future runs that use the threshold based on the model's future data for that RCP scenario –CMIP5_Ff (RCP45) and CMIP5_Ff (RCP85) for the RCP 4.5 and 8.5 scenarios respectively.

There are five combinations of time period (F or H), threshold (f or h), and RCP (4.5 or 8.5). Future heat waves are defined two ways to separate changes due to a general trend (like a regional warming trend) from changes in the heat wave LSMP properties (intensity, frequency, and duration). Comparing 'Ff' to 'Hh' emphasizes changes in LSMP properties. The 'Fh' to 'Hh' comparison includes both the general trend as well as LSMP changes, so how properties differ between 'Fh' and 'Ff' emphasizes the general trend.

2.3. Clustering Methodology

We use two separate calculations 1) to determine which cluster an event belongs to and 2) to assess the strength of each cluster type present in every event.

Previous work, reported in LG2016, showed two groupings of the air parcel trajectories that arrive near the northwest California coast. That location is the center of the hot anomaly at 850 hPa that is fundamental to CCV heat waves. GL2016 choose the hottest 28 heat wave events (from 1977-2010) to form the composite cluster patterns. They show that the ERA-interim (from 1979) and NNRA1 cluster patterns are essentially the same. Here the hottest 32 heat wave events (from 1971-2010) form the composite cluster patterns.

To assign each event to a cluster type in model data, projections are used in a sub-region or 'domain' where there are large and consistent differences between the cluster composites in NNRA1 data and areas that are well above the Earth's surface. After some testing, the domain bounded by135-120W and 40-55N was chosen to determine to which cluster an event belongs. In this domain target values at -2 days lag (i.e. before onset) are used of: temperature anomaly at 850hPa (Ta850) and 500hPa (Ta500) plus zonal wind anomaly at 500hPa (Ua500) .

Projection coefficients (P_{kj}) are calculated for the domain and variables stated above.

$$P_{k,n} = \frac{\sum_i \sum_j (q_{i,j,n} . Q_{k,i,j})}{\sum_i \sum_j (Q_{k,i,j}) 2}$$

Here k indicates cluster type 1 or 2; *n* indicates a date (i.e. during an event) and *i,j* is a grid point in the longitude, latitude domain. The summations are over all grid points in the domain. *q* is the variable during an individual event while *Q* is the corresponding variable in the cluster mean field calculated from the NNRA1 data. There are three combinations of variable, level and time before onset, hence three projection coefficients for each event and cluster type. The three projections are averaged to obtain an average projection for each event and cluster type. The pair of average projections for each event can be drawn on a scatter plot. The larger, positive projection determines the assigned cluster in most events. However, if the average projection onto one cluster differs by less than 0.30 from the average projection on the other cluster or if both projection coefficients are negative, the event is assigned to the 'mixed' category. The method discussed thus far is used only to determine the cluster *type* of an event. The strength of the event is measured with the LSMPi value described next.

2.4. Updating a Large Scale Meteorological Pattern Index (LSMPi)

Grotjahn (2011, 2013, 2016) introduced a "circulation index" (Ci) that measures how similar a pattern on a given day is to the heat wave composite pattern in corresponding variables. The Ci in Grotjahn (2011) uses the temperature anomaly at 850 hPa (Ta850) and meridional wind anomaly at 700 hPa (Va700) values averaged over the event onset dates (labeled 'target composites'). Corresponding daily fields are projected (un-normalized and separately) onto the target composites of Ta850 and Va700 in regions that are highly consistent between ensemble members. The Ci was an optimal weighted combination of these two projections each day. 'Extreme' dates were the hottest 1% of the Tnamax values during the entire data record. The levels and variables were chosen to match the daily climate model data available to the author at that time. Later work, such as GL2016, used different levels, variables, and regions to do the projections and also use more stations in the CCV surface maximum temperature average; again, the choices were dictated by available data and optimized matching.

This study improves upon this Ci definition. To distinguish this new index from the earlier one, it is labeled the LSMP index, or LSMPi. The following approaches are used:

(i). Use only data on the heat wave onset date

(ii). Focus on regions with high consistency (measured by the 'sign count'; Grotjahn, 2011)

(iii). Focus on simple-shaped regions with anomaly extrema (relative maxima and minima) that are also common to both cluster types

(iv). Test spatially-varying weighting proportional to the sign count.

The LSMPi variables and the regions used are these:

Temperature anomaly, Ta at 850 hPa (i.e. Ta850), in region 128-119W, 29-46N

Meridional wind component anomaly, Va at500 hPa (=Va500), in region 142-132W, 37-51N

Zonal wind component anomaly, Ua at 500 hPa (i.e. Ua500), in region 128-111W, 28-37N

A scatter plot can compare the LSMPi values with those CCV-average avTnamax values for all 4880 days of summer (1971-2010) from the NNRA1. Similar plots are in Grotjahn (2013) and Katz and Grotjahn (2014).

The LSMPi was computed using a simple projection of the daily observed field onto the corresponding target composite field over the indicated regions. The match between LSMPi and avTnamax values on dates of extreme avTnamax was improved by including weights in the projection calculation, where the weights, $w_{i,j}$ are proportional to the sign count at each location. Thus, grid points in the region where the anomaly signs are more consistent between past events are given more weight. And grid points with smaller sign count are given less weight when used in the projection calculation. The following equation is used to calculate the LSMP index for the 850hPa temperature.

$$I_{w,n}(T850) = \frac{\sum_i \sum_j w_{i,j} \bar{T}_{\mathrm{a}}(\mathrm{i,j}).\mathrm{T_a}(\mathrm{i,j,n})}{\sum_i \sum_j w_{i,j} \bar{T}_{\mathrm{a}}(\mathrm{i,j})^2}$$

Where: $I_{w,n}(T850)$ is a weighted, normalized projection for a specific day n based on the temperature anomalies at 850hPa level; i and j are the longitude and latitude pointers respectively.

The summations are over the ranges of i and j for the specified regions (above) over which the projection is made. $T_a(i, j, n)$ is the anomaly value of the temperature for that specific day n and grid point (i, j). $\overline{T}_a(i, j)$ is the corresponding target composite at that particular grid point calculated from the onset dates of the 32 events. The weight $w_{i,j}$ is the same as the sign count at that grid point calculated from the 32 onset events. Analogous indices using each velocity component were also calculated from projections over their respective regions defined above.

The weights were adjusted to optimize the LSMPi match for extreme avTnamax values. The circulation index is defined as, $LSMPi = w1*I_{w,n}(T850)+w2*I_{w,n}(V500)+w3*I_{w,n}(U500)$ where V500 (U500) is the 500 hPa meridional (zonal) wind anomaly. Here the w1, w2, and w3 weights are constrained such that w1+w2+w3=1. To optimize the weighting, the root mean square difference between avTnamax and LSMPi for each weight combination of w1, w2, and w3 was calculated. All possible combinations (in 0.01 increments) were tested. An optimal combination (w1=0.68, w2=0.02, w3=0.30) minimized the root mean square difference between the LSMPi value and the avTnamax value over the summers.

These LSMPi values are compared against avTnamax values using scatter plots (shown later). In addition, the distribution of LSMPi values for all days are binned then fit with a curve using the Johnson system (Johnson, 1949) for all days in every group of 40 summers. Estimation of the Johnson parameters is done from quantiles. The procedure of Wheeler (1980) is used. From these fitted curves, we show how the distributions of LSMPi values change between the Hh, CMIP5_Ff, and the CMIP5_Fh cases.

2.5. Determining Extreme Event Skill

This work focuses on extreme events. Hence, some metrics from matching event avTnamax with LSMPi extreme values are calculated:

1. The avTnamax that corresponds to the 95th percentile is called Ts95
2. A cubic polynomial regression line fits only dates when the CCV stations mean (avTnamax) is ≥ Ts95
3. That regression line defines the LSMPi value corresponding to Ts95 and is called LSMPi-Ts95 (LSMPi-Ts95 varies for different combinations of T850, V500, and U500).

Some standard metrics are based on these contingency table quantities:

1. Number of points N_all where either LSMPi ≥ LSMPi-Ts95 *or* the avTnamax is ≥ Ts95
2. Number of points N_s where LSMPi ≥ LSMPi-Ts95 *and* avTnamax is ≥ Ts95 (these are forecast successes)
3. Number of points N_u where avTnamax is ≥ Ts95 and LSMPi < LSMPi-Ts95 (LSMPi is a bust because an event is occurring by this measure but the LSMPi value is below the threshold to signal an event)
4. Number of points N_o where LSMPi ≥ LSMPi-Ts95 and avTnamax is < Ts95 (LSMPi is a bust because it exceeds the threshold to signal an event but the avTnamax values are not

high enough to indicate an event)

The contingency table provides standard indices like FAR (false alarm ratio) and POD (probability of detection). Other researchers have used these indices to detect rare events (Stephenson et al. 2008, Marzban 1998). FAR= N_o/(N_o+N_s) while POD=N_s/(N_u+N_s) It is best if the false alarm ratio is low and the probability of detection is a high value.

2.6. Determining Weighted Model-mean Weights

The models are not equally adept at capturing the number and intensity of heat wave events in the historical period (e.g. GL2016). So, a model-mean should not weight each model simulation equally. Various methods were tested to devise an objective weight for each model's contribution to the weighted model-mean. The Kolmogorov-Smirnov test measuring the distance between cumulative distribution functions (found by the Johnson method) proved unsatisfactory, as some models that matched NNRA1 data less well were ranked better than other models that matched NNRA1 properties better. Several measures of error in Wehner (2013) were tested (with the weight proportional to the inverse of the error) but the weights were similarly unsatisfactory. Since the multi-model average is used to estimate some basic properties of extreme events, such as their intensity, frequency, and distribution of high values, then metrics of those properties are used. The weighting scheme selected uses four, squared, inverse, normalized, model-relative, differences. The difference in variable '*v*' for model '*m*', $d_{v,m}$, is the model value minus NNRA1 value divided by the NNRA1 value of the variable. The inverse of $d_{v,m}$ is used but normalized by the sum of the inverse $d_{v,m}$ values from all models, meaning that the weight is dependent upon the relative corresponding values of other models. Hence, the scheme adapts to the 'competition' by the other models used to compose the multi-model mean. The inverse is defined as $b_{v,m} = (1/d_{v,m}) / \left\{ \sum \left(1/d_{v,l} \right) \right\}$ where the summation is over all the models '*l*', including model '*m*'.

The four variables for each model *m* are: 1) LSMPi mean divided by its standard deviation; 2) the number of days with LSMPi >1 divided by the total number of days; 3) the value of the shape parameter from a generalized Pareto (GP) fit; 4) the value of the scale parameter from the GP fit. These variables are from the 40 year historical period and the weights are assumed to hold for all future periods. The $b_{v,m}$ values for each of the variables are combined to get a root mean squared total, S_m as: $S_m = \sqrt{b_{1,m}^2 + b_{2,m}^2 + b_{3,m}^2 + b_{4,m}^2}$. The final model weight W_m is defined relative to other models by dividing by the sum of the corresponding '*S*' from every model '*l*': $W_m = S_m / \sum S_l$. Therefore, all the W_m values sum to one.

2.7 LSMP pattern metrics

Four metrics are calculated to assess how similar each model's LSMP is to the corresponding reanalysis LSMP.

The LSMP is the ensemble mean of a meteorological field at the onset of all the heat waves in the reanalysis and model 40-year historical periods. Bias ($B_{v,m}$) and percent error ($PE_{v,m}$) for variable 'v' as the temperature anomaly at 850 hPa and model 'm' are:

$$B_{Ta850,m} = \frac{\sum_i^N \sum_j^M \left|w_{i,j}\, C_j\right| \left(MT_{i,j,m} - RT_{i,j}\right)}{\sum_i^N \sum_j^M \left|w_{i,j}\, C_j\right|} \quad , \quad PE_{Ta850,m} = 100. \frac{\sum_i^N \sum_j^M \left|w_{i,j}\, C_j\right| \; \left|MT_{i,j,m} - RT_{i,j}\right|}{\sum_i^N \sum_j^M \left|w_{i,j}\, C_j\, RT_{i,j}\right|}$$

where $1 \le i \le N$ is the range in longitude, $1 \le j \le M$ is the range in latitude, $C_j = \cos(\varphi_j)$ where φ_j is the latitude (in radians) of each grid point, $W_{i,j}$ equals the sign count for the reanalysis ensemble, $RT_{i,j}$ is the value of the *reanalysis* ensemble mean at the point i,j (an average of the 32 events, here), and $MT_{i,j,m}$ is the value of the *model 'm'* ensemble mean at the point i,j (an average of however many events that model 'm' had). The units of $B_{Ta850,m}$ are K. These quantities are used to assess the hot anomaly centered quite close to the area of interest.

Two measures of the structure of the larger portion of the LSMP are the pattern correlation ($Cor_{v,m}$) and reanalysis projection ($Prj_{v,m}$). These quantities are defined for the 850 hPa temperature anomaly as:

$$Cor_{T850,m} = \frac{\sum_{i=istrt}^{iend} \sum_{j=jstrt}^{jend} \left\{\left(MT_{i,j,m} - \overline{MT}_{i,j,m}\right)\left(RT_{i,j} - \overline{RT}_{i,j}\right)\right\}}{\left\{\sum_{i=istrt}^{iend} \sum_{j=jstrt}^{jend} \left(\left(MT_{i,j,m} - \overline{MT}_{i,j,m}\right)^2 \left(RT_{i,j} - \overline{RT}_{i,j}\right)^2\right)\right\}^{\frac{1}{2}}} \, , \, \Pr_{T850,m} = \frac{\sum_{i=istrt}^{iend} \sum_{j=jstrt}^{jend} \left\{MT_{i,j,m}\, RT_{i,j}\right\}}{\sum_{i=istrt}^{iend} \sum_{j=jstrt}^{jend} \left\{RT_{i,j}\right\}^2}$$

The overbar indicates the average value for all the points in the domain. The domain used for this variable is much broader and captures more of the LSMP. For this variable, the domain encompasses the large hot anomaly (centered off the northern California coast) and the cold anomalies flanking it to the west and east. Since the domain includes hot and cold anomalies, the overbar terms tend to be small.

3. Results

3.1. Model Representation of the Primary LSMP

The LSMP that contributes most strongly to the indices is in the temperature anomaly at 850 hPa. Accordingly, how well the models capture this pattern at heat wave onset is a primary indicator of how well the models do in simulating California heat waves. Table 1 lists bias, percent error, pattern correlation, and pattern projection of each model's ensemble mean relative to the ensemble mean of the reanalysis as described in section 2.7.

The bias and percent error are calculated over a small region (128W-119W by 29N-46N) designed to capture the larger and more consistent (as measured by sign count, Grotjahn, 2011) hot anomaly. As discussed in Grotjahn (2011) this anomaly sets up pressure and wind fields to oppose penetration inland of a cooling sea breeze. Many models have a negative bias meaning their temperature anomaly is not hot enough, though two models have a positive bias. The percent error varies from about 10% to nearly 40%. Higher resolution does not guarantee lower bias and percent error.

The pattern correlation and projections extend over a large region (175W-95W by 20N-60N) that captures the stronger pattern of cold-hot-cold anomalies that extends from near the date line to the middle of North America. The pattern correlations range from 0.93 to 0.72 with 8 models having $Cor_{Ta850,m} \geq 0.9$. Hence, the models are doing an excellent job of capturing not just the hot anomaly but the cold anomalies upstream and downstream. (Interested readers can see plots of this LSMP for several models in the supplementary notes.) While the correlation describes the pattern, the projection includes additional information about the magnitude of the anomaly in the model. The projections have a broader range than the correlations. Most models have projection less than one, consistent with their cold bias. Models with larger negative (cold) biases have projections notably less than their correlations. The models with positive (warm) biases have projections that exceed the correlations. Higher resolution only partly yields better pattern match. For example, the bcc models have two quite different resolutions, both models have positive bias, the higher resolution model has larger pattern correlation, but the bias pushes the lower resolution model to a higher projection.

3.2 Past and Future Event Number and Duration

The purpose of this section is to discuss how the climate model heat wave events change between historical and future climate simulations. For this analysis, we compare the CMIP5_Hh and the CMIP5_Fh and Ff cases.

Heatwave definitions include a minimum duration of extremely hot days (Grotjahn, 2011). Figure 1 is a histogram of consecutive days above the threshold (specified in section 2.2). Not surprisingly, longer durations are less common than shorter durations above the threshold. For the CMIP5_Hh case, almost all the higher resolution models (Figure 1a) do a reasonable job simulating the distribution found in the reanalysis data. Most of the coarser resolution models generally tend to overestimate the duration of events (Figure 1b).

Not surprisingly, Figure 1 shows that heat wave durations increase in the future simulations when using each model's historical threshold (Fh cases), and more so for RCP 8.5 data. For example, in the HADGEM2-CC model RCP8.5 Fh scenario, heat wave events that last for 5 days are more common than heat wave events lasting 3 or 4 days and there are three times as many events as in the model's Hh data. Inmcm4 and NorESM1-M have large numbers of events in the CMIP5_Fh cases, but these models also have a many more events in their CMIP5_Hh data than are present in the reanalysis. Other models have between three and four times as many extreme heat wave events in Fh versus Hh data. Table 2 lists the total number of events for each scenario by each model as well as the weighted model mean.

Most models examined have increased average duration. In the CCSM4 model, the RCP8.5_Fh events are, on average, 2.6 days longer than for Hh simulations while the RCP4.5_Fh events are 1.3 days longer; these averages are over 6 ensemble runs in each case. For the bcc-csm1-1-m model, the increase of average duration is 0.5 days in the RCP4.5 Fh case, but 2.5 days in the RCP8.5 Fh case. HadGEM2-CC has a much larger change in average duration: 2.3 days for RCP4.5 and 5.9 days for RCP8.5. A few models (notably the MIROC models) show much longer increases in average event duration.

In comparing the Hh and Ff cases, there is generally little change in the average duration or the general shape of the histograms, especially for models having more than one ensemble member (CCSM4 with 6 ensemble members; HadGEM2-CC with 3 ensemble members). Hence, the frequency and duration of the weather patterns, i.e. the LSMPs producing the heat waves are likely little-changed from their historical values. This point is developed further below.

The longest events are in general between 7-10 days in the higher resolution models in Hh simulations. For all models the longest event becomes longer in each future simulation, typically doubling (or more) in length for RCP4.5 Fh cases and tripling (or more) for RCP8.5 Fh cases. The longest day has increased from Hh to RCP8.5 Fh by 20 days in CCSM4 and to 45 days in HadGEM2-CC. Comparing RCP8.5 Ff to RCP8.5 Fh the longest day has increased by about 18 days in CCSM4 and 41 days in HadGEM2-CC. The longest day between Hh and RCP8.5_Ff increases in 8 out of the 13 models. The longest day has increased more than three times in CCSM4 and more than six times in the HadGEM2-CC for the CMIP5_Fh RCP8.5 scenario. However the increase in the CMIP5_Fh for the RCP4.5 scenario is about two times for CCSM4 and 3 times for HadGEM2-CC. However, comparing Hh and Ff cases finds little difference in the length of the longest events (similar to the average duration results).

Similar histograms are shown in Grotjahn (2016) who uses durations above one standard deviation for Hh and Fh simulations by CCSM4. The threshold he used is lower than the threshold used here. He found RCP8.5 durations to be most common at four and five days, ahistogram structure different than found for CCSM4 here, but similar to the result for HadGEM2-CC. He also found the number of events declines more slowly for longer durations than shown here. His results are consistent with the a general warming comparable to one standard deviation, but much less than the 95th percentile used here.

A June-September climatology shows a linear type increase of the zonal wind and temperature anomalies in the projection domain between the RCP4.5 and the RCP8.5 simulations. But, there is not a clear increase in the events from the RCP4.5 to RCP8.5 simulations. Somee RCP8.5 CMIP5_Fh simulations (CCSM4, bcc-csm1-1-m, CNRM-CM5, and inmcm4, GFDL-ESM2G and GFDL-ESM2M) do increase the number of events from the RCP4.5 to the RCP8.5 simulations. However, the other models (including coarser resolution models MIROC-ESM, MIROC-ESM-CHEM and FGOALS-g2) have fewer heat events in RCP8.5 than RCP4.5, contrary to one's expectation.

The multi-model weighted average number of events is also included in Table 2. The numbers of events are essentially the same between Hh and Ff simulations (35.6 and 36.3 respectively) and are similar the reanalysis number of 32. However, the number of events using historical thresholds in the future (Fh data) is four times as large for RCP8.5 simulations. The average duration in the multi-model average is 4.19 d for Hh, 4.35 d for RCP8.5_Ff, and 8.73 d for RCP8.5_Fh calculations. Again, the Hh and Ff values are similar to the reanalysis number (4 d) but using historical thresholds, the duration is more than twice as long on average.

3.3. Past and Future Number of Events by Cluster Type

Most heat wave events have LSMPs that cluster into one of two types. However, a few events are not clearly of either type and are designated as 'mixed' type using the projection methodology described in section 2.3. The average projection values for each pair of cluster types for each event are shown as scatter plots in Figure 2. The projection method was developed for the NNRA1 data (which match corresponding values for ERA-Interim data as a check). The NNRA1 data in Figure 2 nicely separate events along a line between the two clusters, with one mixed type. The NNRA1 data show that if an event projects strongly on one cluster type, then that event projects weakly or negatively the other cluster type. Although simulated historical heat waves in the models are not so neatly along a line, most model events separate into one of the two types in a way that is similar to the reanalysis result. As noted in Table 2, the models vary a bit in terms of their relative fractions of type 1, 2, or mixed. The models do tend to have more mixed events, but the proportion of events in each type is not much different than the renanalysis for most models.

The projection procedure was applied to the RCP4.5 and RCP8.5 simulations using historical thresholds (Fh). These data are not plotted but the numbers of events of each type are included in Table 2 for the RCP8.5 simulations. The greater number of events in the future using historical thresholds is not evenly split between the two cluster types but is disproportionately found in type 2. Cluster type 2 is characterized by a preexisting hot anomaly in southwestern Canada, but the future climatology in the models is several degrees warmer than historically, especially over the contents and extending over the adjacent oceanic areas. (Interested readers can see the CCSM4 future climatology in the supplemental materials.) The domain used for the cluster type designation has a cool anomaly for type 1 and a warm anomaly for type 2 at 850 and 500 hPa. Hence, the future climatology alone favors the type 2 projection.

As noted, the future climatology (Ff) has a similar number of events as in the historical period. The split between the two types changes between historical (Hh) and future (Ff) simulations in the models. In the CCSM4 and MIROC-ESM-CHEM models type 2 events double and type 1 are fewer. In contrast, the CNRM-CM5 model has half as many more type 1 but fewer type 2 events. Other models change the balance between event types between these extremes. The balance between the two event types in RCP4.5 simulations are similar though some models have opposite changes compared to RCP8.5 results. The models do not show a systematic change. Thus, the multi-model average in the future (Ff) is very similar to the recent past (Hh). In short, neither cluster type LSMP is more common in the future.

The Hh, Fh, and Ff results taken together indicate that the amount of variability is not obviously changed but that the increase in heat waves (based on historical thresholds) is due primarily to a change in the climatology, i.e. to the 'global warming signal'.

As discussed above, the number of events does not consistently increase from RCP4.5 to RCP8.5 Fh simulations. This is also the case for the number of events in both cluster types when comparing Ff thresholds. When comparing Fh thresholds most models have an increase of type 2 between RCP4.5 to the RCP8.5 simulations; the exceptions are: HadGEM2-CC, GFDL-CM3, and the MIROC models.

3.4. Past and Future Cluster Strength

The strength of each event is measured by the largest avTnamax that occurs during the event. These largest avTnamax values can be further stratified by the cluster type. Figure 3 shows the evolution of event strength by cluster type over each 40-year period. As above, cases with anomalies defined using historical climate means are designated Hh and Fh, while anomalies defined from future climate means are labeled Ff.

The large future increase of cluster type 2 events in the Fh results is immediately obvious in the preponderance of blue symbols. The increased strength of events and the increased number of cluster 2 events in the future (using historical means) are both easily seen.. In general, most models tend to have very similar scatter in the Hh and Ff panels. But, within the Ff panels, the number of events per decade increases towards the end of the period for most models, especially for RCP8.5. The HadGEM2-CC model's historical preference for cluster type 1 trends towards more type 2 events in the RCP8.5_Ff panel. However, CNRM-CM5 maintains its preference for type 1 events in Ff panels.

Since avTnamax values are normalized by the standard deviation, the peak values of those future temperatures in some models are quite high. For RCP8.5_Fh, CCSM has a half-dozen events exceeding four standard deviations above the historical mean. Similar results are found for other models, including the other four highest resolution models, plus NorESM1-M and GFDL-CM3. The other two GFDL models and FGOALS-g2 do not have quite as strong events. The remaining models, especially the MIROC models have stunningly high peak average temperatures as numerous events exceed 5 standard deviations and in the MIROC-ESM model two events exceed eight standard deviations above the historical mean. The MIROC models and to a lesser degree the bcc-csm1-1 results are consistently different from the other models in having larger scatter and extreme avTnamax values in historical as well as future climatological situations.

3.5. Past and Future LSMP Index Distributions

Each panel in Figure 4 compares the LSMP index (LSMPi) with the extreme values of the CCV-average surface temperature (avTnamax) at the corresponding time. These panels are similar to Figure 4 in Grotjahn (2013) and Figure 1 in Katz and Grotjahn (2014). Contingency table scores: false alarm ratio (FAR) and the probability of detection (POD) are included in each panel. FAR and POD are defined as in 'binary' weather forecasting. FAR is the number of 'false alarms' divided by the sum of the 'hits' plus false alarms. A 'hit' is where the avTnamax values are above the 95% threshold and the LSMPi is above the regression curve value for that avTnamax threshold, i.e. both quantities indicate a heat wave. A 'false alarm' is where the LSMPi value is above its threshold but the avTnamax is not. The POD is the number of hits divided by the sum of the hits plus misses. A miss is where the avTnamax is above its threshold but the LSMPi is not. A better match between LSMPi and avTnamax is when FAR is smaller and POD is larger. FAR and POD both range from 0 to 1.

The LSMPi was developed to best fit avTnamax on the few onset dates of heat waves using the NCEP-NCAR reanalysis data. The climate models also have a strong correspondence between high LSMPi and high avTnamax. Nearly all models outperform the reanalysis judging from the FAR and POD values. Collapsing the relationship to a regression curve (Figure 4) shows that the relationship between LSMPi and avTnamax varies between models. Most models have a nearly linear regression curve meaning the match between LSMPi and avTnamax extends from moderate to high values of avTnamax. Such models that show a consistent LSMPi to avTnamx relationship for very high temperatures reinforce applying LSMPi to future climate simulations. However, MIROC-ESM models have a large spread of low LSMPi values during high avTnamax dates while the inmcm4 model has a large range of avTnamax values for high LSMPi dates, both situations reduce the match between the two quantities; but since both situations do not occur together in these models, their FAR and POD scores are better than for the reanalysis..

Figure 5 shows the historical and future distributions of LSMPi>1 values. This figure is similar to Figure 7 in GL 2016, but the figure here shows all the extreme values not just LSMPi values on the onset days. The NCEP-NCAR reanalysis distribution is plotted in every panel as a blue dotted curve. The historical simulations (dotted red curves) seem to underestimate the standard deviation of the LSMPi distribution in several models, especially CCSM4, NorESM1-M, the MIROC models, and FGOALS-g2. However, the bcc models, CNRM-CM5, and HadGEM2-CC values match the reanalysis very well on this high tail of the distribution.

Figure 5 shows future scenarios using both historical (Fh) and future (Ff) climatologies to define anomalies. As mentioned above, the number of events and relative strength of the events are very similar between the Ff and Hh results. Ff and Hh distributions in Figure 5 also have highly similar high tails, though some models differ from this general conclusion. HadGEM2-CC and the GFDL models have notably less probability density values in Ff than in Hh results for both RCP scenarios. Model inmcm4 has less density for RCP8.5 than either historical or RCP4.5 results. Since there are many more heat waves that last longer in the future when using historical thresholds, the Fh curves in Figure 5 are systematically shifted to higher values relative to the Hh and Ff curves. Superficially, the RCP8.5 and RCP4.5 distributions (dashed line curves) appear to be approximately parallel with the historical curves. The RCP8.5 Fh curve is less steep for the CNRM-CM5 and GFDL-CM3 model results. The RCP8.5 Fh curve is steeper for the bcc-csm1-1-m, inmcm4, NorESM1-M, and GFDL-ESM2M models. Grotjahn (2016) found an increasingly negative skew during the 21st Century for an index similar to the LSMPi applied to CCSM4 output; but this is harder to see in Figure 4 because so little of the LSMPi range is shown.

Some qualitative impressions from Figure 5 can be made quantitative by looking at the scale and shape parameters from a Generalized Pareto distribution (GP) fit to the data shown in Figure 5. The GP scale parameter (Figure 6a) varies by ~0.1 between models relative to the multi-model mean and the reanalysis value (0.32). The direction of the change in GP scale between cases is generally consistent. Except for the CNRM-CM5 and inmcm4 models: the scale increases for RCP4.5_Fh and even more for RCP8.5_Fh. The amount of increase varies greatly between models. However, the multi-model average is a third larger for RCP4.5 and more than half again larger for RCP8.5 Fh. The GP shape parameter is negative for the reanalysis and nearly all cases by all the models. Negative shape means the tail is unbounded. The models are not consistent about the change of GP shape between the cases. Because the shape results are so broad that parameter is not shown (but shape is plotted in the supplemental materials for an interested reader).

Return value also provides information on the distribution's high tail and is shown in Figure 6b. The 20-year return value may be interpreted as that value having a 5 % chance of being exceeded in any particular year. The return values of CMIP5_Ff cases are generally very close to the Hh historical values for each model (LSMPi = 1.3-2). The differences between Fh and Hh values fall within the error bars and are smaller than the range among the models. So again, the large scale pattern for the heat wave is not occurring more intensely in the future if one uses the future climatology to define the anomalies. The return values for Fh cases are systematically >50% larger than the historical values (LSMPi = 2-2.8). The multi-model averages are 1.76 for Hh, 2.14 and 2.24 for RCP4.5 and 8.5, respectively. As Figure 4 shows, different models have a different relation between LSMPi value and corresponding near surface temperature. For many models LSMPi increases more slowly than temperature; so, LSMPi 20-year return values >2 imply very high if not unprecedented surface temperatures.

A broad estimate of the surface temperatures that correspond to 50 highest avTnamax values in each model is shown in Figure 6c. The estimate is calculated as follows. The average of the 50 highest avTnamax values is found for each case and model. Each average is multiplied by delT = 3.97K. This delT is the value used to normalize the temperature anomalies on average for the CCV stations during summer. The difference between this future climate value and the historical value for the model is plotted in Figure 6c. Relative to future climatology (blue and red dots), the models vary about zero, consistent with other results shown above. The future simulations relative to historical values finds a consistent increase that is larger for RCP8.5_Fh. The amount of increase ranges from 2 to 8K for RCP4.5_Fh and 4 to 11K for RCP8.5. The multi-model averages are: 3.3 and 6 for RCP4.5 and 8.5, respectively.

Inspection of Figure 3 has qualitative evidence for a trend of increasing number of events within each time period for several models and especially the Hh and RCP8.5 Ff groupings. A simple quantitative metric for such a trend (Figure 6d) is to subtract the average of the 30 highest avTnamax values in the first 20 years from the corresponding average over the last 20 years of each period. There is no clear trend in the RCP4.5 data, but most models do have increasing avTnamax in their historical and RCP8.5 data. Grotjahn (2016) showed similar results for CCSM4 data and slightly different comparison periods. The multi-model average trends are 1 for Hh, and 1.4 (1.5) for RCP8.5 Ff (Fh).

4. Summary

This report focuses on how general properties of CCV heat wave events change for two future scenarios simulated by 13 climate models. Future climate results are shown using anomalies defined relative to either historical climatology ('Fh' data) or the climatology of the future period ('Ff' data). The future scenarios are RCP4.5 and RCP8.5. The future simulations by each model are compared against the historical simulations ('Hh' data) by the model to detect relative changes. LG2016 discovered two types of CCV heat wave patterns leading up to onset while GL2016 showed that these climate models develop both types. There are thus five such groupings of model output between one Hh, and two scenarios each of Fh and Ff data. Each of these five categories can be further split into the two cluster types. NCEP-NCAR reanalysis data for a period of the same (40-year) length are used for comparison.

The heat wave type and intensity can be related to the upper air large scale meteorological patterns (LSMPs) as shown in our prior publications. Previous work (LG2016) showed that these data yield LSMPs that are essentially the same as those for two other reanalyses. This work improves upon the the LSMP-based schemes: on heat wave intensity (GL2016) and on heat wave type (LG2016). As demonstrated in these prior works, the use of an index like the LSMPi provides a compact and accurate way of characterizing the larger scale weather pattern developed by a model during a heat wave. To make surface maximum temperatures intercomparable, they are normalized by the local standard deviation and averaged over the CCV; the result is labelled 'avTnamax'. A strong relationship exists between the daily LSMPi and avTnamax values. The link is obvious in scatter plots and a corresponding regression curve is calculated for each model. The connection between

LSMPi and avTnamax is *stronger* in the models than it is in the reanalysis. So, properties of the LSMPi characterize heat events in the models and it is useful to examine the statistics of such indices. Furthermore, how well each model's historical simulations match four statistical properties of the reanalysis LSMPi defines weights used to calculate multi-model means. The models that match the reanalysis better are given more weight in the multi-model mean.

Similar to GL2016, most models capture the frequency of heat wave events, though some models develop twice as many heat waves. The distributions of duration are comparable to that in the reanalysis, though the models developing many more heat waves have a larger fraction of short (3-day) events. The split between event types varies between models, as noted in GL2016. In the future scenarios, when historical thresholds and climatology are used, there are many more events and their durations are much longer than in corresponding historical simulations. In terms of the cluster types, the majority of the increased events are cluster type 2 which has a pre-existing heat wave over Canada not present in cluster type 1. (An interested reader can find the pattern in LG2016 and also in the supplementary materials along with the change in future climatology for CCSM4.) In the future scenarios, the models have higher average temperatures over the continents, thereby explaining the asymmetric preference for type 2 heat waves. However, when the heat waves are defined as extremes relative to the future climatology, then the number of events and the proportion of each cluster type both are very similar to the corresponding historical values. Therefore the large scale patterns that create the heat waves are not occurring more or less frequently in the future. To the extent the measure of the LSMP (here called the LSMPi) represents the variability of the summer temperatures (as shown by Grotjahn, 2011) then the future variability is the same as in the historical simulations. That result means that the increase in heat waves and their intensity is primarily due to a warming of the average conditions. These general results are seen in all the other metrics shown.

Other metrics include tracking the avTnamax values in models. Again, Hh and Ff data are very similar, while Fh data have many more days and higher values. How high the values reach varies greatly between models. Most models have peak avTnamax values between 2-3 standard deviations for Hh and Ff calculations while most Fh values range between 2-4 (2-5) standard deviations above the mean in RCP4.5 (RCP8.5) data. However, a few models have Fh values up to 8 standard deviations.

The number of events is larger for RCP4.5 than RCP8.5 in six models and vice versa for the other seven. This result is not counter-intuitive since those six models have events with much longer durations than historically. Fewer events can occur in a set period of time when they last longer. For example, in the HadGEM2-CC simulations, the number of Fh events in RCP4.5 is greater than in RCP8.5, the longest event is more than twice as long (53 vs 22 days), and the average duration increases from 6.63 to 9.94 days in the RCP8.5 Fh data. Longer duration events are not consistent across the models in Ff data. Some models, like CCSM4 have slightly longer average duration in Ff than Hh data, while other models like bcc-csm1-1-m show slightly shorter average duration. So as with other results, the patterns are not lasting longer than corresponding historical patterns when the future climatology is used to define them.

The change in climatology shows up as a trend in the RCP8.5 data, but not in the RCP4.5 data. Grotjahn (2016) found a similar result for the CCSM4 model; here, there is variation between the other 12 models and no consistent trend for the RCP4.5 data *within the 2061-2100 period*. The extreme values in RCP4.5 data are consistently larger than historical values in all models, though the amount of increase varies widely, by a factor of three from the 1961-2000 values and those a century later. The RCP8.5 data increase even more and vary by nearly a factor of three, as well, for this suite of models.

The extreme value statistics for heat waves confirm equivalent behavior between Hh and Ff data and a shift of distributions for Fh data. Multi-model averages of the Generalized Pareto scale parameter for LSMPi in Fh data show an increase(by more than 50%) and an increase of the 20-year return period LSMPi value by almost 30% in the RCP8.5 data, both are consistent with the shift of the distribution to higher values. Extreme temperatures also increase. An estimate based on historical scaling finds the multi-model average is >3C warmer for RCP4.5 and 6C hotter for RCP8.5 scenarios compared to historical conditions.

Acknowledgements

This research was funded in part by NSF grant 1236681 and also supported by the USDA National Institute of Food and Agriculture, Hatch project Accession #1010971. Coding assistance was provided by Dr. Yun-Young Lee. Additional support provided by NASA grant NNX16AG62G and Department of Energy Office of Science award number DE-SC0016605, ``An Integrated Evaluation of the Simulated Hydroclimate System of the Continental US.". NCEP Reanalysis data are provided by the NOAA/OAR/ESRL PSD, Boulder, Colorado, USA, from their Web site at http://www.esrl.noaa.gov/psd/data/gridded/data.ncep.reanalysis.html. ECMWF ERA-Interim data used in this project are from the ECMWF data server: http://apps.ecmwf.int/datasets/data/interim-full-daily/ . We acknowledge the World Climate Research Programme's Working Group on Coupled Modelling, which is responsible for CMIP, and we thank the climate modeling groups (listed in Table 1 of this paper) for producing and making available their model output. For CMIP the U.S. Department of Energy's Program for Climate Model Diagnosis and Intercomparison provides coordinating support and led development of software infrastructure in partnership with the Global Organization for Earth System Science Portals. These model data are available from the Earth System Grid Federation at https://esgf-data.dkrz.de/search/cmip5-dkrz/ . Data plotted here are available from the lead author and from http://grotjahn.ucdavis.edu/EWEs .

References

Coumou, D., Robinson, A., & Rahmstorf, S. (2013). Global increase in record-breaking monthly-mean temperatures. *Climatic Change*. *118* 771–82. https://doi.org/10.1007/s10584-012-0668-1

Dee, D., Uppala, S., Simmons, A., Berrisford, P., Poli, P., Kobayashi, S., et al. (2011). The ERA‐Interim reanalysis: Configuration and performance of the data assimilation system, *Quarterly Journal of the Royal Meteorological Society*, *137*(656), 553-597. https://doi.org/10.1002/qj.828

Grotjahn, R. (2011). Identifying extreme hottest days from large scale upper air data: a pilot scheme to find California Central Valley summertime maximum surface temperatures, *Climate Dynamics*, *37*(3-4), 587-604. https://doi.org/10.1007_s00382-011-0999-z

Grotjahn, R. (2013). Ability of CCSM4 to simulate California extreme heat conditions from evaluating simulations of the associated large scale upper air pattern, *Climate Dynamics*, *41*(5-6), 1187-1197. https://doi.org/10.1007/s00382-013-1668-1

Grotjahn, R. (2016). Western North American extreme heat, associated large scale synoptic-dynamics, and performance by a climate model, in J. Li, R. Swinbank, R. Grotjahn, H. Volkert (Eds.), *Dynamics and Predictability of Large-scale, High-Impact Weather and Climate Events*, (pp. 198–209). Cambridge, England: Cambridge University Press,

Grotjahn, R., Black, R., Leung, R., Wehner, M.F., Barlow, M., Bosilovich, M., et al. (2016). North American extreme temperature events and related large scale meteorological patterns: a review of statistical methods, dynamics, modeling, and trends. *Climate Dynamics, 46*, 1151–1184. https://doi.org/10.1007/s00382-015-2638-6

Grotjahn, R. & Lee, Y.-Y. (2016). On climate model simulations of the large-scale meteorology associated with California heat waves, *Journal of Geophysical Research: Atmospheres*, *121*, 18–32. https://doi.org/10.1002/2015JD024191

Horton, R.M., Mankin, J.S., Lesk, C., Coffel, E., & Raymond, C.. (2016), A Review of Recent Advances in Research on Extreme Heat Events. *Current Climate Change Reports*, *2*, 242-259. https://doi.org/10.1007/s40641-016-0042-x.

Johnson, N.L. (1949). Systems of frequency curves generated by methods of translation. *Biometrika*, *36*. 149-176.

Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven D., Gandin, L., et al. (1996). The NCEP/NCAR 40-year reanalysis project, *Bulletin of the American Meteorological Society*, *77*(3), 437-471.

Katz, R.W., & Grotjahn, R., (2014). "Statistical methods for relating temperature extremes to Large-Scale Meteorological Patterns." US CLIVAR *Variations*, *12*, 4-7.

Lee, Y-Y. & Grotjahn, R. (2016). California Central Valley summer heat waves form two ways. *Journal of Climate, 29*, 1201-1217. https://doi.org/10.1175/JCLI-D-15-0270.1.

Lehmann, J., Coumou, D., Frieler, K., Eliseev, A.V., & Levermann, A..(2014). Future changes in extratropical storm tracks and baroclinicity under climate change. *Environmental Research Letters 9*(8)084002. https://doi.org/10.1088/1748-9326/9/8/084002

Marzban, C. (1998). Scalar measures of performance in rare-event situations. *Weather and Forecasting, 13*, 753–763.

Perkins, S.E, Alexander, L.V., & Nairn, J.R. (2012). Increasing frequency, intensity and duration of observed global heatwaves and warm spells. *Geophysical Research Letters, 39* (20), L20714, http://dx.doi.org/10.1029/2012GL053361

Petoukhov ,V., Rahmstorf, S., Petri, S., & Schellnhuber, H.J. (2013). Quasiresonant amplification of planetary waves and recent Northern Hemisphere weather extremes. *Proceedings of the National Academy of Sciences of the United States of America, 110*, 5336–5341. https://doi.org/10.1073/pnas.1222000110.

Russo, S., Dosio, A., Graversen, R.G., Sillmann, J., Carrao, H., & Dunbar, M.B (2014). Magnitude of extreme heat waves in present climate and their projection in a warming world. *Journal of Geophysical Research: Atmospheres, 119*, 500–512. https://doi.org/10.1002/2014JD022098

Screen, J.A, & Simmonds, I. (2014). Amplified mid-latitude planetary waves favour particular regional weather extremes. *Nature Climate Change, 4*,704–709. https://doi.org/10.1038/NCLIMATE227

Stephenson D.B, Casati, B., Ferro, C.A.T., Wilson, C.A. (2008), The extreme dependency score: a non-vanishing measure for forecasts of rare events. *Meteorological Applications, 15*(1), 41–50, https://doi.org/10.1002/met.53

Teng, H., Branstator, G., Wang, H., Meehl, G.A., & Washington, W.M. (2013). Probability of US heat waves affected by a subseasonal planetary wave pattern. *Nature Geoscience 6*, 1–6. https://doi.org/10.1038/ngeo1988

Wheeler, R.E. (1980). Quantile estimators of Johnson curve parameters. *Biometrika, 67*(3) 725-728 https://doi.org/10.2307/2335153

Wehner, M.F. (2013). Very extreme seasonal precipitation in the NARCCAP ensemble: model performance and projections. *Climate Dynamics, 40*, 59-80. https://doi.org/10.1007/s00382-012-1393-1

Figure Captions

Figure1. Histogram of heat waves duration (in consecutive days) for CMIP5 models for each of the groupings: Hh, Ff and Fh (both RCP4.5 and RCP8.5 scenarios). The historical period is 1961-2000 while the future period is 2061-2100. Included in the figure are the length of the longest event and the average duration. For models with more than one ensemble member, each bin is divided by the ensemble size. The longest event in each ensemble member was found, added together, and then divided by the number of ensembles for that model to produce the number shown. a) Six CMIP5 models with corresponding NCEP-NCAR reanalysis values for 1971-2010 shown for comparison. b) seven more CMIP5 models.

Figure 2. Projection coefficients onto each cluster type for all heat waves in the reanalysis and the models. The projections are onto upper air variables in a specific region as detailed in the text. Red dots are events that are primarily type 1 while blue dots are primarily type 2; green dots are mixed

type events. These data are for 40-year historical periods. Events in all ensemble members are shown; CNRM-CM5, NorESM1-M, MIROC-ESM, both bcc, and all three GFDL models have three ensemble members; HadGEM2-CC and FGOALS-g2 have two members, and the remainder one member.

Figure 3. Maximum avTnamax temperature during each event as a function of time in each 40-year period. The peak value of each event is color-coded such that red circles are cluster type 1, blue circles designate type 2, and green circles are the mixed type. The layout of the reanalysis and model groupings matches figure 1: a) reanalysis and six models; b) seven more models. To make the results in different models and groupings comparable, only one ensemble member is used for each grouping.

Figure 4. Scatterplots of daily avTnamax (abscissa) and corresponding LSMP index (ordinate) for every day of the CMIP5_Hh simulations. The best fit curve uses the points where avTnamax is >1. Also included are the FAR (False alarm ratio) and the POD (probability of detection).

Figure 5. Distribution functions of LSMPi >1 for all historical (Hh) summer days (June-September). The NCEP-NCAR reanalysis (1971-2010) (blue dotted) curve is on all panels for reference. Model data are shown in a format similar to Figure 2._Red dotted curves are model Hh (1961-2000) data. Future scenarios (2061-2100) use green curves for RCP 4.5 and purple curves for RCP 8.5 data, with solid lines for Ff data and dashed lines for Fh data.

Figure 6. Distribution properties for the models. The black dots are Hh data, the red dots are RCP4.5_Ff data, the blue dots are RCP8.5_Ff data, the green dots are RCP4.5_Fh, and the purple dots are RCP8.5_Fh data. Corresponding values for the multi-model weighted average and the NCEP-NCAR reanalysis is also shown. a) Generalized Pareto (GP) scale parameter for the extremes in the models examined for the five groupings. The threshold for the extremes is LSMPi>1 (The LSMPi values >1 were all declustered to make the data independent prior to the calculation as recommended for GP calculations. To calculate the GP function, we have used all the ensembles available for each model. b) 20-year return values of LSMPi in the models and reanalysis. c) Temperature anomaly difference from the Hh data of the 4 groups of future scenarios. The anomaly in each group is the mean of the 50 largest avTnamax values for each group multiplied by the delT value, where detT is the magnitude of the temperature normalization averaged over the summer and all CCV stations. Here the delT value equals 3.97C. d) The trend within each grouping, calculated as the average of the 30 largest avTnamax values during the last 20 years minus the corresponding values for the first 20 years. These values are also multiplied by the delT value, so these trends have units of C/20 years.

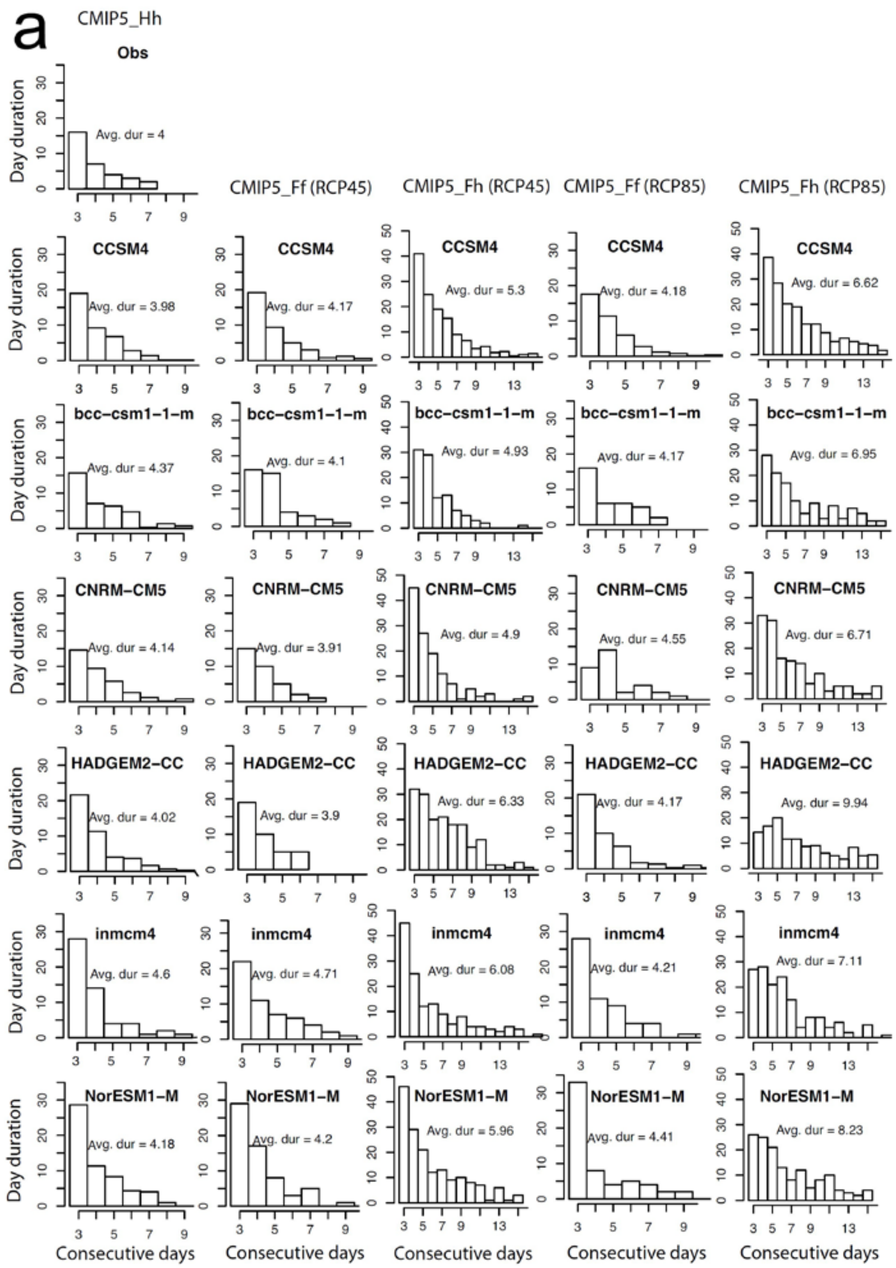
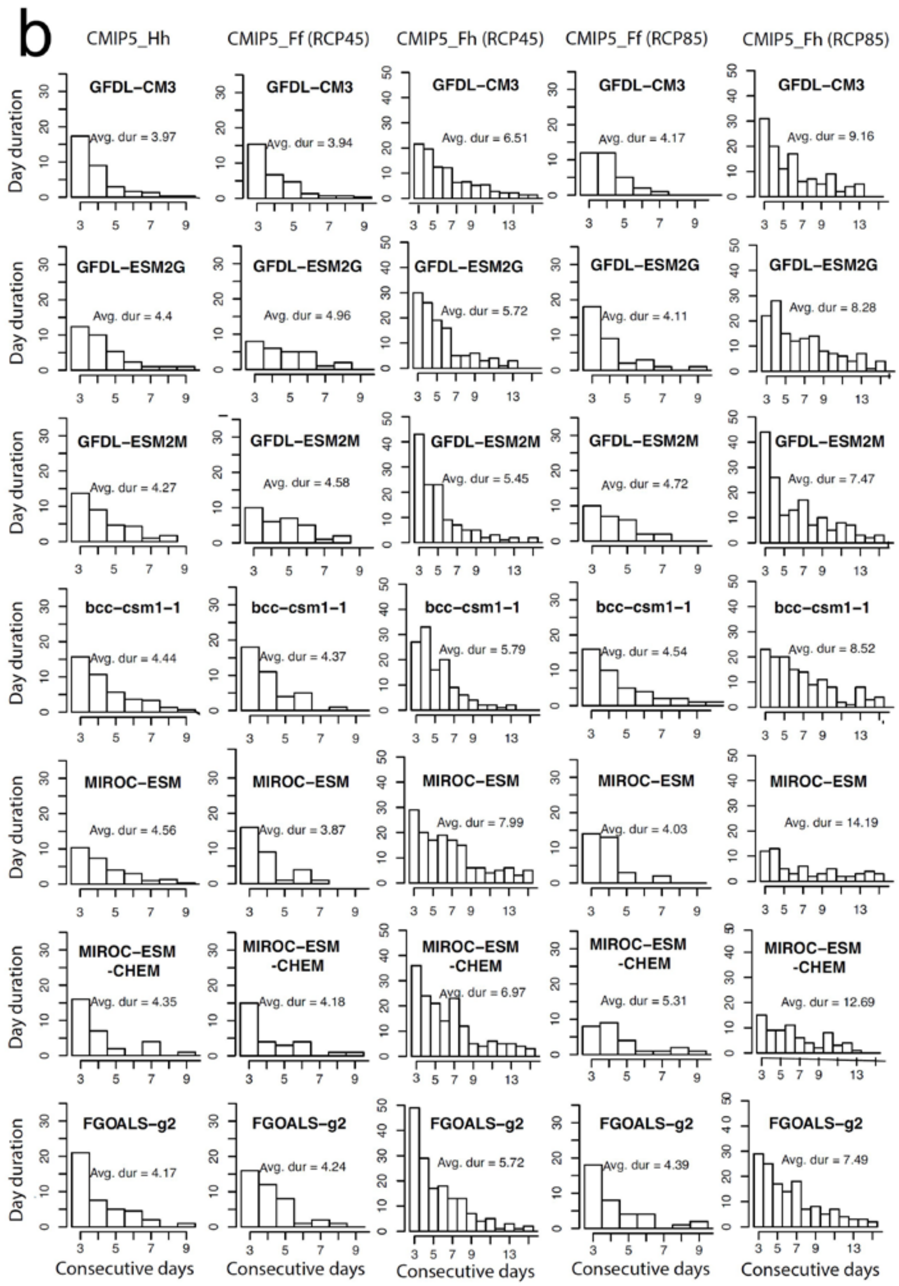


Figure1. Histogram of heat waves duration (in consecutive days) for CMIP5 models for each of the groupings: Hh, Ff and Fh (both RCP4.5 and RCP8.5 scenarios). The historical period is 1961-2000 while the future period is 2061-2100. Included in the figure are the length of the longest event and the average duration. For models with more than one ensemble member, each bin is divided by the ensemble size. The longest event in each ensemble member was found, added together, and then divided by the number of ensembles for that model to produce the number shown. a) Six CMIP5 models with corresponding NCEP-NCAR reanalysis values for 1971-2010 shown for comparison. b) seven more CMIP5 models.


b
CMIP5_Hh
CMIP5_Ff (RCP45)
CMIP5_Fh (RCP45)
CMIP5_Ff (RCP85)
CMIP5_Fh (RCP85)
GFDL–CM3
Avg. dur = 3.97
Avg. dur = 3.94
Avg. dur = 6.51
Avg. dur = 4.17
Avg. dur = 9.16
GFDL–ESM2G
Avg. dur = 4.4
Avg. dur = 4.96
Avg. dur = 5.72
Avg. dur = 4.11
Avg. dur = 8.28
GFDL–ESM2M
Avg. dur = 4.27
Avg. dur = 4.58
Avg. dur = 5.45
Avg. dur = 4.72
Avg. dur = 7.47
bcc–csm1–1
Avg. dur = 4.44
Avg. dur = 4.37
Avg. dur = 5.79
Avg. dur = 4.54
Avg. dur = 8.52
MIROC–ESM
Avg. dur = 4.56
Avg. dur = 3.87
Avg. dur = 7.99
Avg. dur = 4.03
Avg. dur = 14.19
MIROC–ESM -CHEM
Avg. dur = 4.35
Avg. dur = 4.18
Avg. dur = 6.97
Avg. dur = 5.31
Avg. dur = 12.69
FGOALS–g2
Avg. dur = 4.17
Avg. dur = 4.24
Avg. dur = 5.72
Avg. dur = 4.39
Avg. dur = 7.49
Day duration
Consecutive days

.

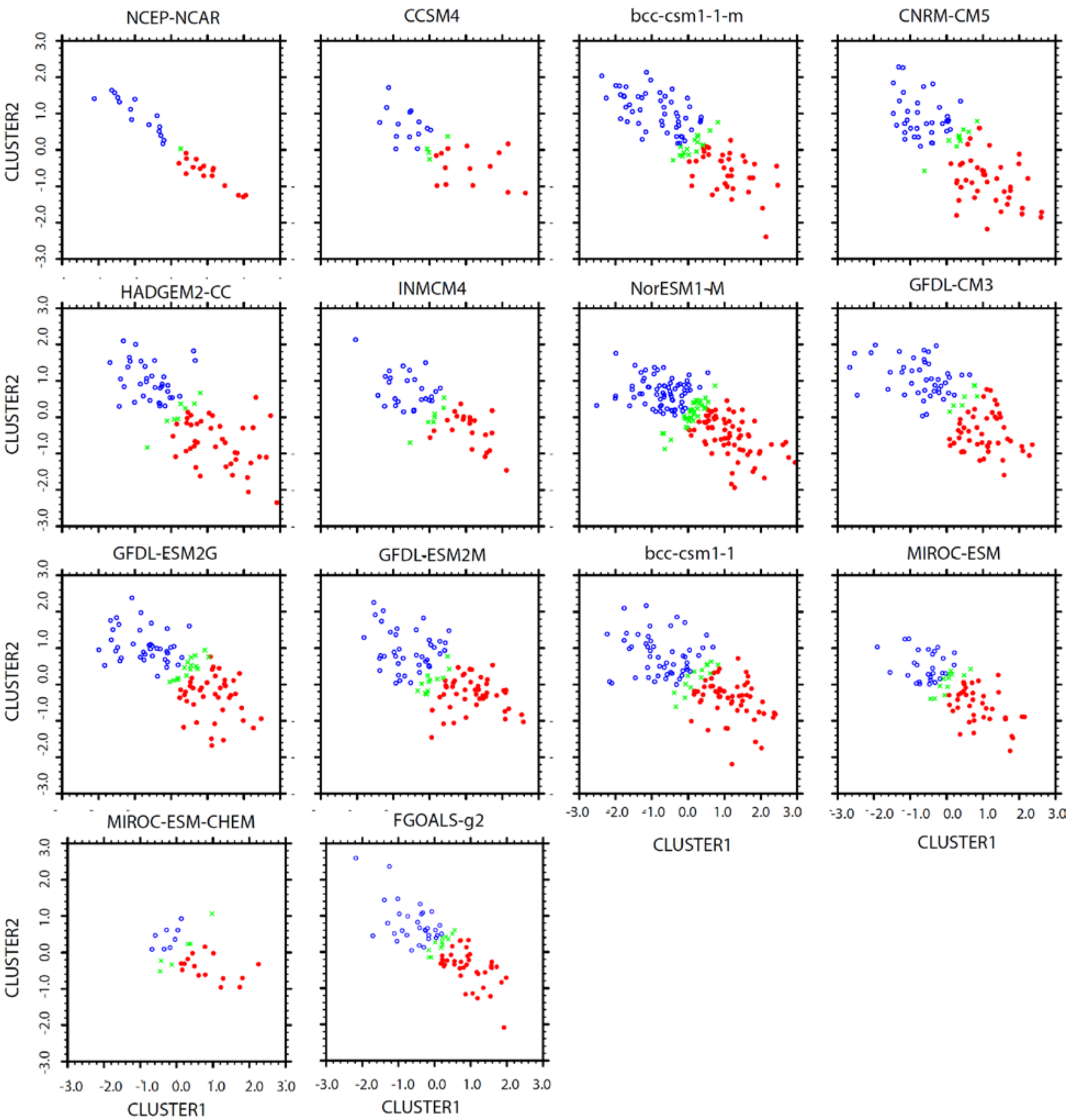


Figure 2. Projection coefficients onto each cluster type for all heat waves in the reanalysis and the models. The projections are onto upper air variables in a specific region as detailed in the text. Red dots are events that are primarily type 1 while blue dots are primarily type 2; green dots are mixed type events. These data are for 40-year historical periods. Events in all ensemble members are shown; CNRM-CM5, NorESM1-M, MIROC-ESM, both bcc, and all three GFDL models have three ensemble members; HadGEM2-CC and FGOALS-g2 have two members, and the remainder one member.

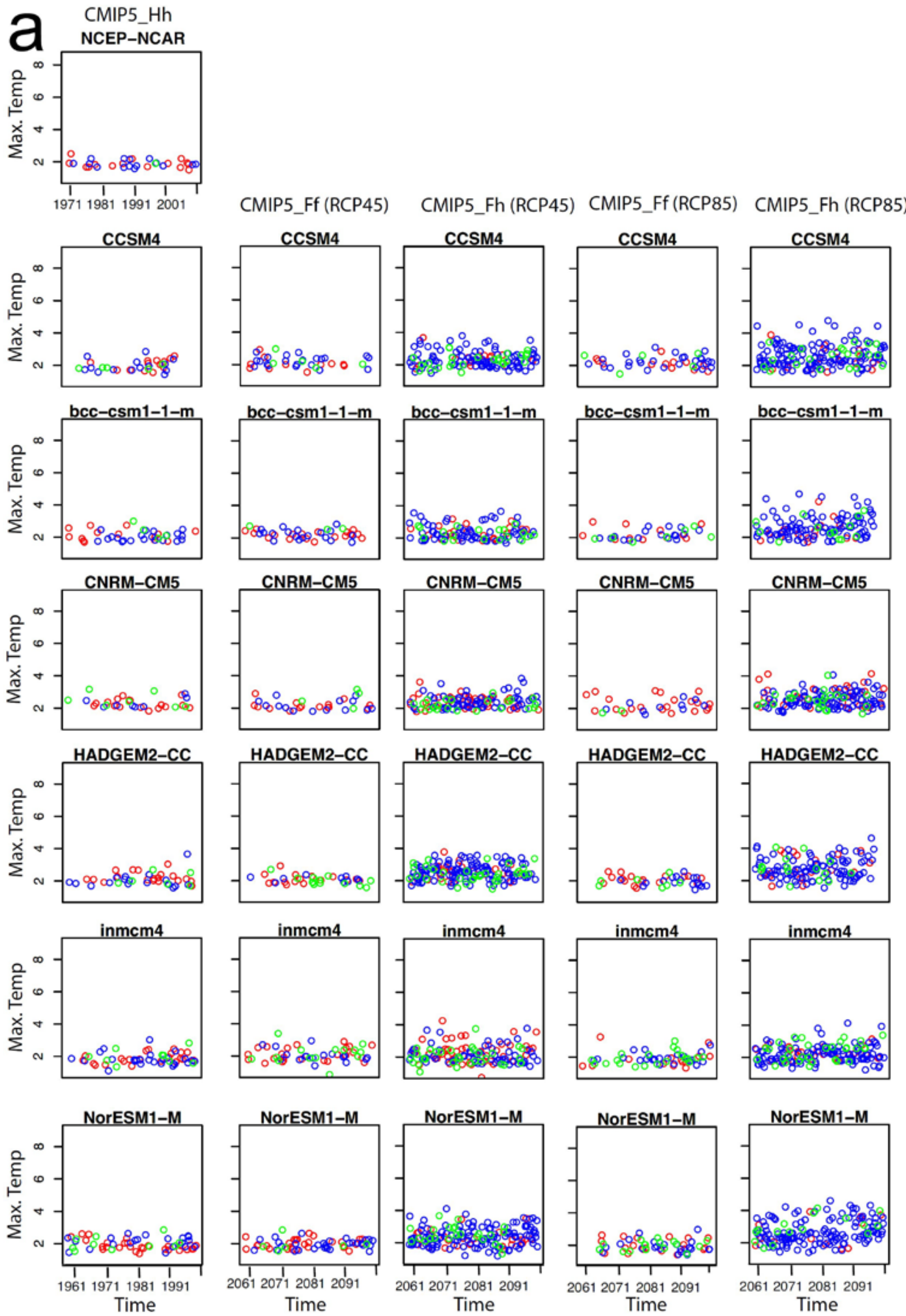
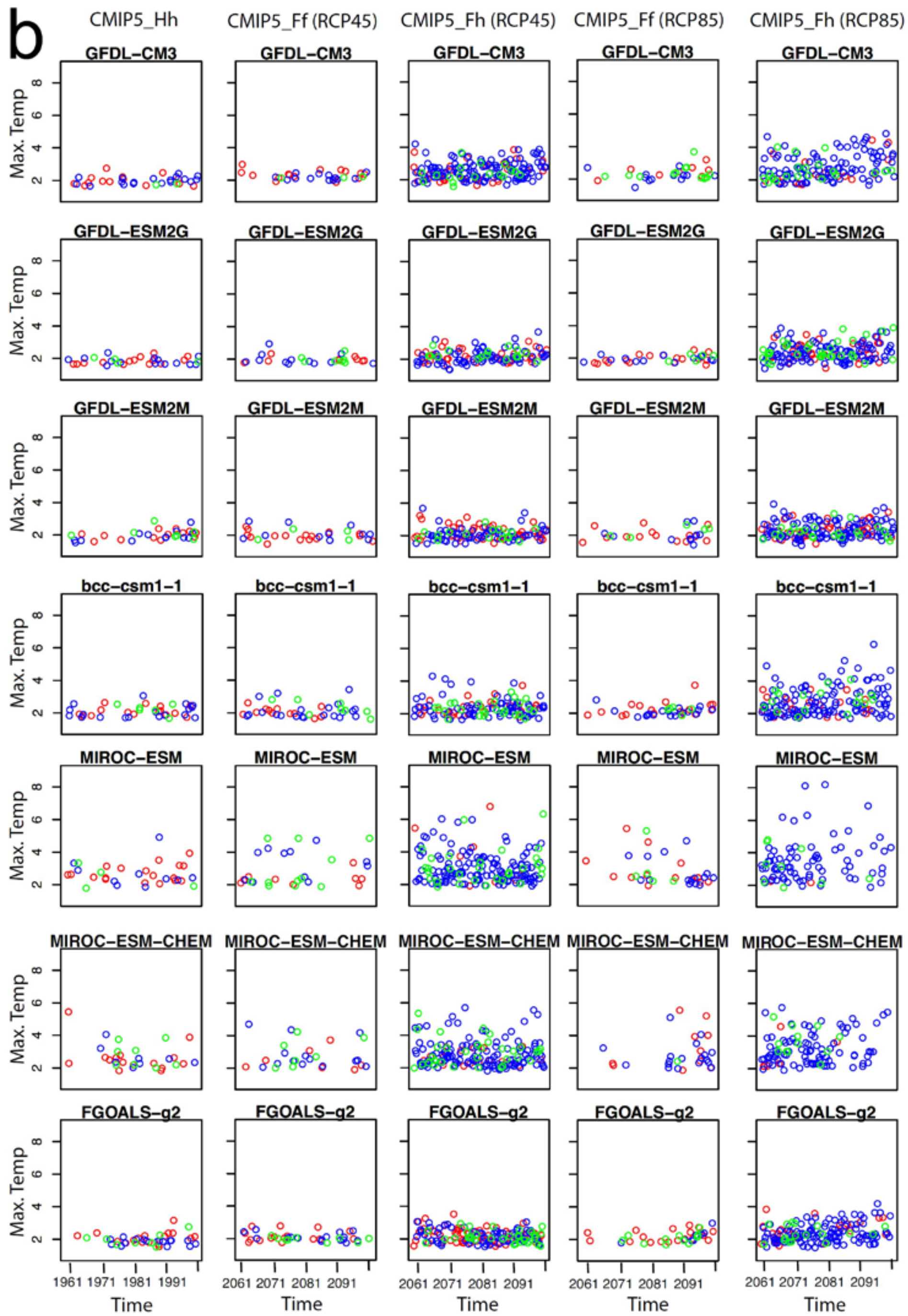


Figure 3. Maximum avTnamax temperature during each event as a function of time in each 40-year period. The peak value of each event is color-coded such that red circles are cluster type 1, blue circles designate type 2, and green circles are the mixed type. The layout of the reanalysis and model groupings matches figure 1: a) reanalysis and six models; b) seven more models. To make the results in different models and groupings comparable, only one ensemble member is used for each grouping.


b
CMIP5_Hh
CMIP5_Ff (RCP45)
CMIP5_Fh (RCP45)
CMIP5_Ff (RCP85)
CMIP5_Fh (RCP85)
GFDL-CM3
GFDL-ESM2G
GFDL-ESM2M
bcc-csm1-1
MIROC-ESM
MIROC-ESM-CHEM
FGOALS-g2
Max. Temp
Time
1961 1971 1981 1991
2061 2071 2081 2091

927
928

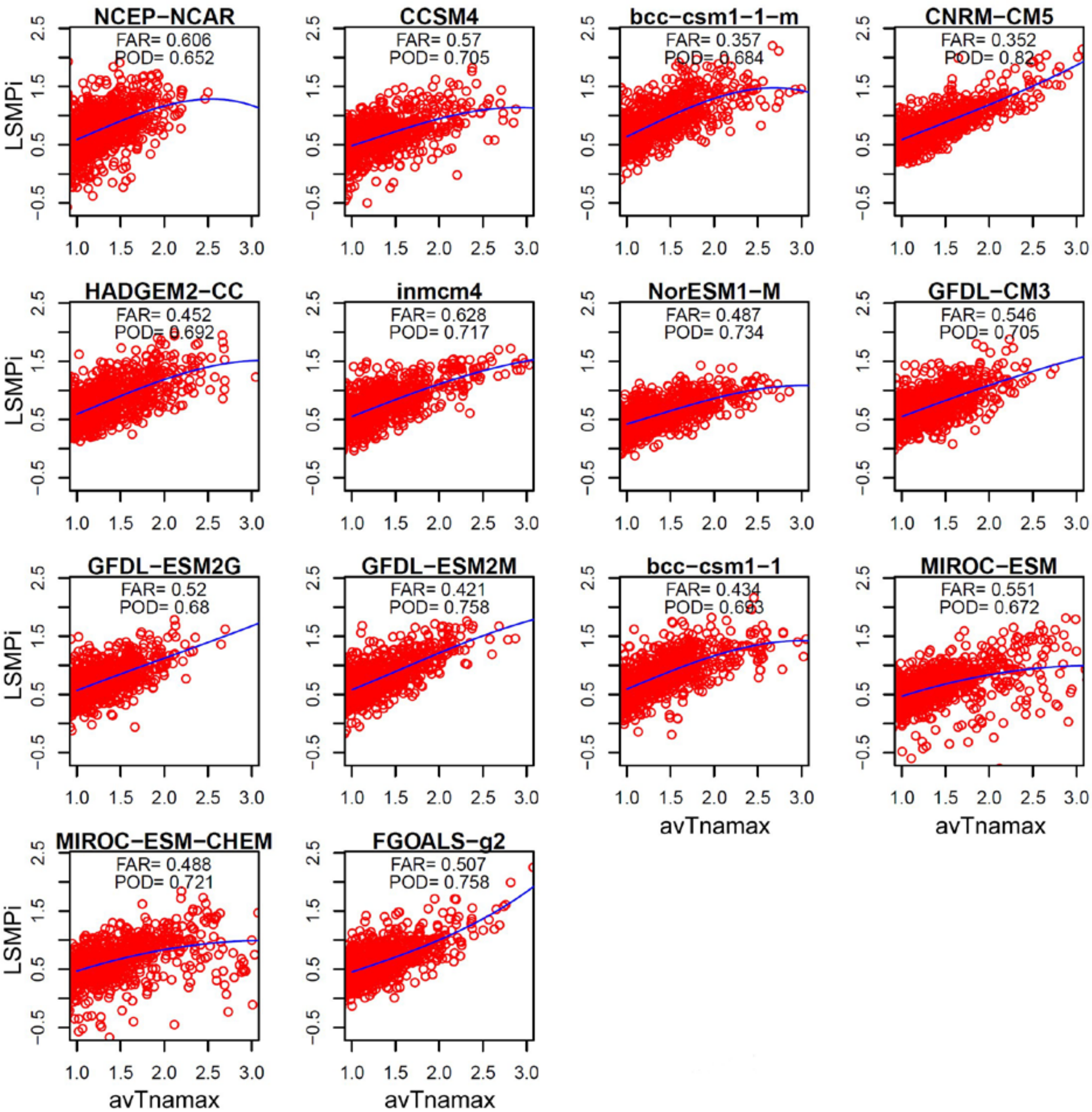


Figure 4. Scatterplots of daily avTnamax (abscissa) and corresponding LSMP index (ordinate) for every day of the CMIP5_Hh simulations. (For models with ensembles archived, only one ensemble run is shown.)The best fit curve uses the points where avTnamax is >1. Also included are the FAR (False alarm ratio) and the POD (probability of detection).

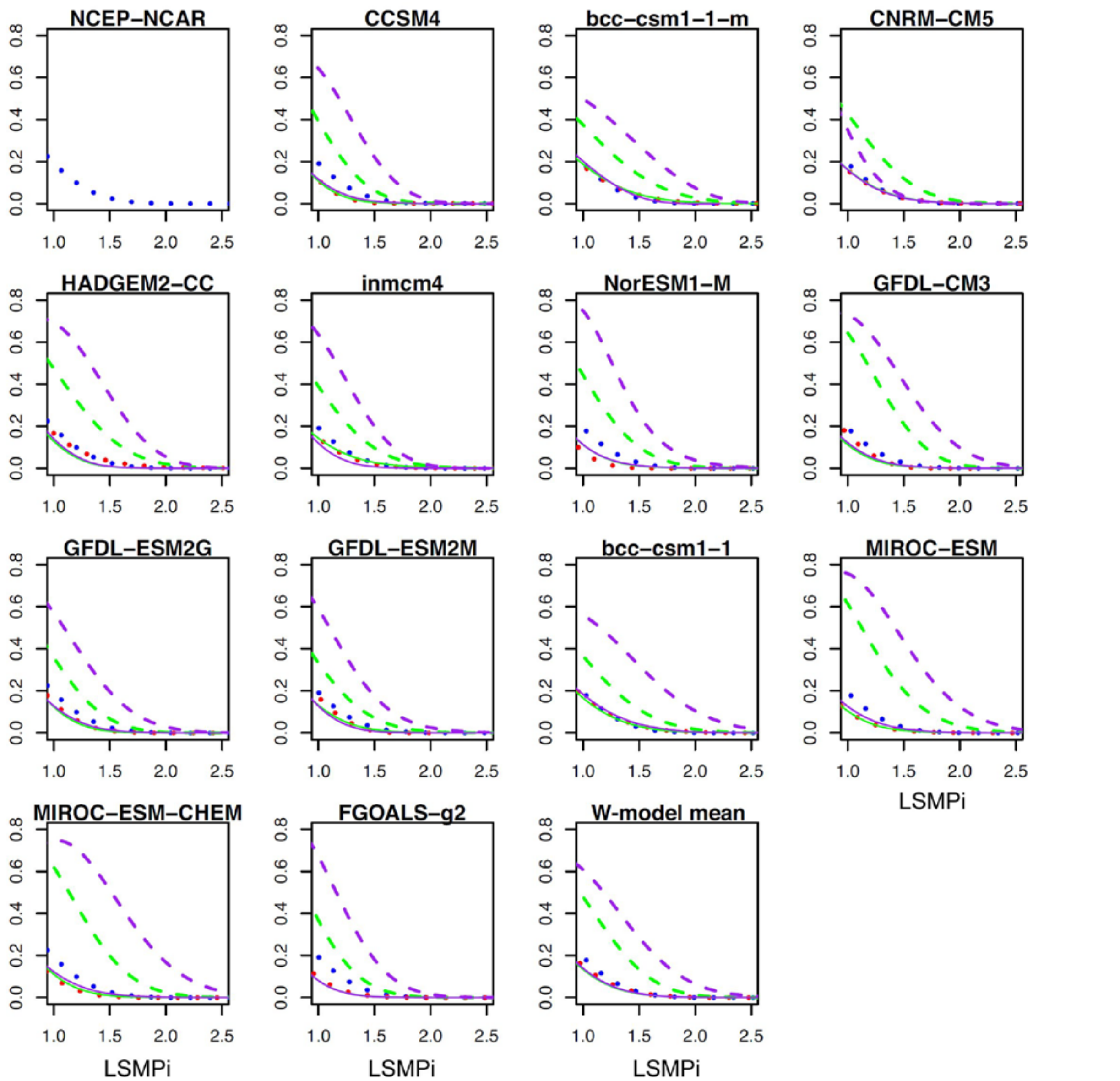


Figure 5. Distribution functions of LSMPi >1 for all historical (Hh) summer days (June-September). The NCEP-NCAR reanalysis (1971-2010) (blue dotted) curve is on all panels for reference. Model data are shown in a format similar to Figure 2._Red dotted curves are model Hh (1961-2000) data. Future scenarios (2061-2100) use green curves for RCP 4.5 and purple curves for RCP 8.5 data, with solid lines for Ff data and dashed lines for Fh data.

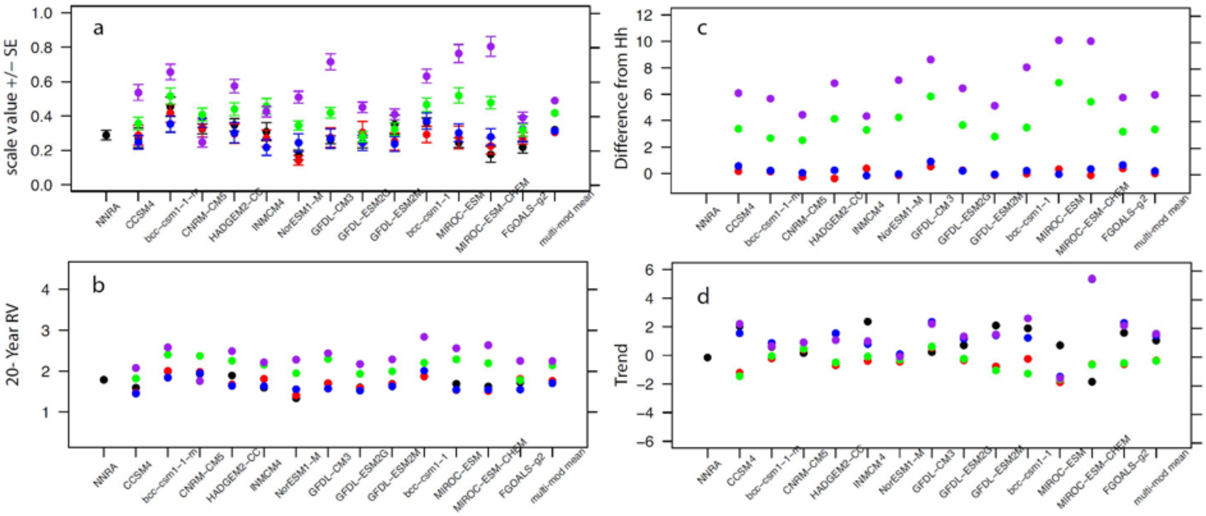


Figure 6. Distribution properties for the models. The black dots are Hh data, the red dots are RCP4.5_Ff data, the blue dots are RCP8.5_Ff data, the green dots are RCP4.5_Fh, and the purple dots are RCP8.5_Fh data. Corresponding values for the multi-model weighted average and the NCEP-NCAR reanalysis is also shown. a) Generalized Pareto (GP) scale parameter for the extremes in the models examined for the five groupings. The threshold for the extremes is LSMPi>1 (The LSMPi values >1 were all declustered to make the data independent prior to the calculation as recommended for GP calculations. To calculate the GP function, we have used all the ensembles available for each model. b) 20-year return values of LSMPi in the models and reanalysis. c) Temperature anomaly difference from the Hh data of the 4 groups of future scenarios. The anomaly in each group is the mean of the 50 largest avTnamax values for each group multiplied by the delT value, where detT is the magnitude of the temperature normalization averaged over the summer and all CCV stations. Here the delT value equals 3.97C. d) The trend within each grouping, calculated as the average of the 30 largest avTnamax values during the last 20 years minus the corresponding values for the first 20 years. These values are also multiplied by the delT value, so these trends have units of C/20 years.

Table 1. *Metrics of model ability to capture the LSMP anomaly temperature at 850 hPa.*

Model	Ta_{850} Bias (K)	Ta_{850} Error (%)	Pattern correlation	Projection (95-175W; 20-60N)	Horizontal Resolution (lon x lat)
CCSM4	-0.20	9.8	0.93	0.91	288x192
Bcc-csm1-1-m	0.92	19.2	0.92	1.21	320x160
CNRM-CM5	-0.43	15.6	0.83	0.81	256x128
HADGEM2-CC	-0.24	10.3	0.90	0.85	192x144
INMCM4	-1.11	23.1	0.90	0.70	180x120
NORESM1-M	-1.92	38.5	0.84	0.60	144x96
GFDL-CM3	-0.37	14.5	0.91	0.90	144x90
GFDL-ESM2G	-0.52	16.6	0.92	0.95	144x90
GFDL-ESM2M	-0.15	11.8	0.89	0.83	144x90
BCC-CSM1-1	0.19	17.9	0.91	0.98	128x64
MIROC-ESM	-1.58	34.7	0.85	0.56	128x64
MIROC-ESM-CHEM	-1.35	33.3	0.72	0.54	128x64
FGOALS-G2	-1.05	25.3	0.90	0.71	128x64

Table 2: *Number of events occurring during 40 year periods for historical (hh; 1961-2000) and future climate (fh; 2061-2100) scenarios for models and the multi-model weights and means. Reanalysis data from 1971-2010 included for comparison.*

Model	CMIP5_hh			CMIP5_fh(RCP8.5)			CMIP5_ff (RCP8.5)			W_m
Event types	# event	Cluster 1	Cluster 2	# event	Cluster 1	Cluster 2	# event	Cluster 1	Cluster 2	
NCEP-NCAR	32	16	15							
CCSM4	34	15	14	168	17	128	44	12	27	.1109
bcc-csm1-1-m	36.67	13.33	17	126	16	97	41	17	19	.0534
CNRM-CM5	33.33	13.67	12.67	154	33	98	33	21	10	.0935
HadGEM2-	44	20.5	17.5	136	22	99	41	17	17	.0947

CC										
inmcm4	58	23	26	166	28	107	58	14	23	.0168
NorESM1-M	58.67	23	23	162	14	131	59	19	24	.0125
GFDL-CM3	33.33	16	14.33	143	18	103	33	8	12	.2076
GFDL-ESM2G	33.67	14	13	167	35	100	35	17	13	.1059
GFDL-ESM2M	34.33	15.33	13.33	171	43	106	29	14	10	.1047
bcc-csm1-1	41.33	18.67	17.33	159	21	121	41	17	19	.0754
MIROC-ESM	28	12.67	9.67	92	2	81	33	13	15	.0578
MIROC-ESM-CHEM	31	15	8	110	6	92	29	10	18	.0595
FGOALS-g2	41.5	19.5	15.5	161	28	115	38	19	8	.0072
Multi-model weighted average	35.6	15.8	14.2	147.7	22.8	104.5	36.3	14.0	15.6	